Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 89 (2014) 1080 – 1088

Procedia Engineering

www.elsevier.com/locate/procedia

16th Conference on Water Distribution System Analysis, WDSA 2014

Analytical Leakages Localization in Water Distribution Networks Through Spectral Clustering and Support Vector MACHINES. The Icewater Approach

A. Candelieri[a,b]*, D. Soldi[b], D. Conti[a], F. Archetti[a,b]

[a]*Consorzio Milano Ricerche, via Roberto Cozzi 53, Milano 20125, Italy*
[b]*Department of Computer Science, Systems and Communication, University of Milano-Bicocca,viale Sarca 336, Milano 20126, Italy*

Abstract

An approach based on hydraulic simulation and machine learning is presented, aimed at improving leakage management via analytical leak localization and reducing time and costs for investigation and rehabilitation of the Water Distribution Network. Hydraulic simulation is used to run different leakage scenarios by introducing a leak on each pipe, in turn, and varying its severity. The approach has been validated on two WDNs: a Pressure Management Zone in Milan (Italy) and a District Metered Area in Timisoara (Romania), the two pilots of the EU-FP7-ICT project ICeWater, obtaining a high reliability (>90%) in localizing a wide set of simulated leaks.

Peer-review under responsibility of the Organizing Committee of WDSA 2014
Keywords: leakage management, leak localization, spectral clustering, support vector machine, simulation

1. Introduction

Nowadays urban Water Distribution Networks (WDNs) suffer leakage, mainly due to the age of the infrastructures, implying failures, large amounts of Non Revenue Water (NRW) and high costs for energy (i.e., pumping) and rehabilitation, while budgetary constraints are becoming more strong. The International Water Association (IWA) highlighted the relevance to improve the leakage management process [1], which is usually divided in three consecutive steps [2]: assessment, detection and physical localization. Several studies proposed to improve localization

* Corresponding.
E-mail address: candelieri@milanoricerche.it

Peer-review under responsibility of the Organizing Committee of WDSA 2014
doi:10.1016/j.proeng.2014.11.228

through the analysis of data collected by computer-based systems usually adopted in WDNs, such as Supervisory Control And Data Acquisition (SCADA), Automatic Metering Readers (AMR), GIS and hydraulic simulation software. Many approaches use machine learning, statistics, probabilistic modeling have been investigated [3-6], with the common idea that actual modifications in flow and pressure within the WDN are linked to a set of leaky pipes: hydraulic simulation software may be therefore used to simulate a wide set of leaks, store variations and then discover the inverse relation between variations and leaky pipes [7-9]. While most of the proposed approaches try to localize a leak on pipes, in [10] a combination between hydraulic simulation and classification learning has been developed to identify leaks on junctions.

This paper presents an analytical framework that uses: 1) extensive simulation of leaks for data generation, 2) network-based Spectral Clustering to group together leaks implying similar variations in pressure and flow, 3) classification learning (i.e., Support Vector Machine, SVM) to discover the relation linking variations in pressure and flow to a limited set of probably leaky pipes (i.e., a cluster of those provided by Spectral Clustering). The approach also proposes a strategy to support cost-effective placement of flow and pressure meters, identifying the best trade-off between reliability in localization and deployment costs. All the results are related to a real test case, a Pressure Management Zone (PMZ) of the WDN in Milan, Italy, one of the two pilots of the FP7-ICT project ICeWater co-funded by the European Commission.

2. Method and Materials

2.1. Description of the pilot and the data generation process

In the following Fig.1, the two ICeWater pilots are depicted: the Pressure Management Zone (PMZ) “Abbiategrasso” in Milan, Italy, and the District Metered Area (DMA) “Neptun” in Timisoara, Romania. Both the pictures have been obtained by using EPANET, a hydraulic simulation software widely used for modeling WDNs and downloadable for free from the Environmental Protection Agency web site (http://www.epa.gov/nrmrl/wswrd/dw/epanet.html). EPANET permits to perform what-if simulation and it can be also integrated with optimization models for supporting decisions both at operational, planning and strategic level.

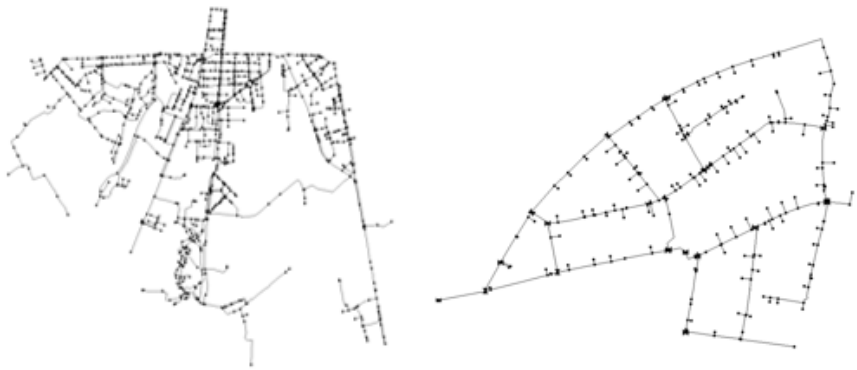

Fig. 1. The PMZ “Abbiategrasso”, in Milan, Italy (left) and DMA “Neptun”, in Timisoara, Romania (right)

Abbiategrasso is a PMZ consisting of 1212 junctions (612 are consumption points) and 1385 pipes; the overall pipe infrastructure is long about 116905m. Pipes length ranges from 0.25m to 844.92m (average 84.41m), pipes diameter ranges from 50mm to 900mm (average 244.92mm). Neptune is a DMA consisting of 335 junctions (92 are consumption points) and 339 pipes; the overall infrastructure is long about 4000m. Pipes length ranges from 0.034m to 83.808m (average 12.778m), pipes diameter ranges from 50mm to 500mm (average106.917mm).

2.2. Clustering Leakage Scenarios and Quality Measure

In the proposed approach, EPANET is used to simulate a wide set of leakage scenarios, consisting in placing, in turn, a leak on each pipe and varying its severity in a given range. At the end of each run, EPANET provides pressure and flow values at each junction and pipe, respectively. Then, variations in pressure and flow, induced by the leak, are computed with respect to the simulation of the faultless network. Results obtained are stored in a dataset, together with the information related to the affected pipe and the damage severity. Each row of this dataset can be represented as a vector with the following notation:

$$v_i = \left\{ l_j, c_k, f_1, \ldots, f_{n_P}, p_1, \ldots, p_{n_N} \right\} \tag{1}$$

where l_j is the link with the leakage, c_k is the leakage severity, $f_1, \ldots, f_{nP}$ are the flow variation measured over the n_P links of the network and $p_1, \ldots, p_{nN}$ are the pressure variation measured over the n_N nodes of the network. Starting from these vectors, $\tilde{v}_i$ are defined removing the labels concerning the pipe (l_j) and the leakage severity (c_k) from the vectors, leaving only the pressure and flow variations. More details about the leakage scenarios generation process as well as the pressure-dependent leak modeling have been firstly described in [7]. Clustering leakage scenarios previously generated through EPANET is the core of the proposed approach. The aim is to group together scenarios (rows of the dataset) that are similar in terms of variations in pressure and flow induced by the corresponding leak. Only information on pressure and flow (that is the effect of the leak) is taken into account during this process, while information on leaky pipe and leak severity is ignored. Several clustering algorithms are available; all of them need a specific measure (distance or similarity) to be defined in order to compare two objects, that in this case are two vectors of pressure and flow variations at junctions and pipes. At the end of clustering process, a measure should be adopted to enable an evaluation of the quality of the solution with respect to the goal. Although several measures have been proposed to evaluate the validity of clustering procedures, evaluating the capability to localize leak has required the definition of a specific measure, namely "Localization Index", already proposed in a previous work of the authors. The Localization Index for each cluster (LI_k) requires to retrieve the information on leaky pipe related to each scenario and is then computed as the number of distinct pipes related to the scenarios in that cluster with respect to the overall number of pipes in the WDN:

$$LI_k = \frac{n_P - n_P^k}{n_P - 1} \tag{2}$$

where n_P is the overall number of pipes of the WDN and n_P^k is the number of leaky pipes of the scenarios into cluster k (with $k \in \{1, \ldots, K\}$, K the overall number of clusters).

The maximum value of LI_k is $LI_k = 1$ that is obtained when the cluster k contains scenarios all related to leaks simulated only on one pipe (i.e., $n_P^k = 1$). On the other hand, the minimum value of LI_k is $LI_k = 0$ that is obtained if the cluster k contains scenarios referred to all the pipes of the WDN (i.e., $n_P^k = n_P$).

While in the previous work of the authors the overall localization index (LI) has been computed as the simple average of LI_k, in this case the average has been weighted by the number of distinct pipes in each cluster:

$$LI = \frac{\sum_{k=1}^{K} LI_k \cdot n_P^k}{\sum_{k=1}^{K} n_P^k} \tag{3}$$

where K is the overall number of cluster. In this work another relevant measure is proposed to evaluate how much a clustering algorithm is able to put in the same cluster scenarios related to same (leaky) pipe and to all the different severity values. This index has been named "Quality of Localization" and is defined, for each cluster k, as:

$$QL_k = \frac{\sum_{j \in P} \frac{n_{s_j}^k}{|S|}}{n_P^k} \tag{4}$$

where S is the set of different severity used (and $|S|$ is the overall number of severity values used), P is the set of the pipes of the network, n_{sj}^k is the number of scenarios in cluster k associated to the *j-th* pipe, and n_P^k is the number of distinct pipes related to the scenarios in cluster k. The maximum value of QL_k is $QL_k = 1$ that is obtained when in the cluster k are all the scenarios related to all the severity values of the correspondent leaky pipes. The overall Quality of Localization for a clustering algorithms is given by the average of QL_k.

Finally, a global index LI^* is defined, combining Localization Index and Quality of Localization

$$LI^* = LI \times QL \tag{5}$$

Amongst the available clustering algorithms, in this paper were used both traditional algorithms (*K*-Means and Farthest-First) and graph-based algorithm (Spectral Clustering).

2.3. Identifying leaky pipes through Support Vector Machines

After clustering the leakage scenarios, the next step consists in discovering a reliable relation between the variations in pressure and flow, due to a leak, and the correspondent cluster, which permits to retrieve the set of correspondent leaky pipes. This relation allows reducing time for investigations and rehabilitation: when a leak is detected (e.g., with traditional methods, such as Minimum Night Flow analysis [3]), the actual pressure and flow measurements at the monitoring points are compared with those obtained through simulation of the faultless network model, in order to compute the variations in pressure and flow and finally identify only a restricted number of pipes to physically check. In their previous study, the authors proposed to compare the obtained vector of values with those related to the centroids of the clusters, and selected the cluster associated to the most similar centroid.

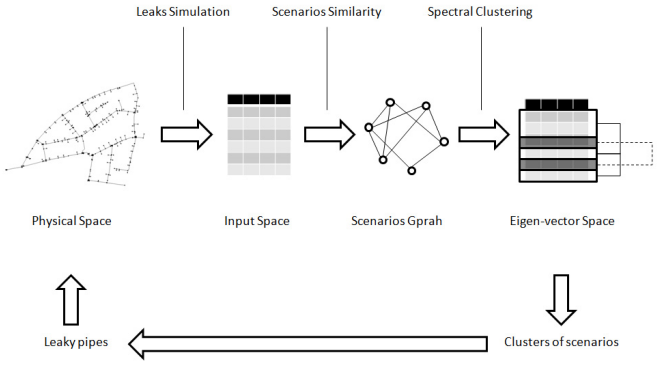

Fig. 2. Overall approach proposed: Spectral versus traditional clustering of leakage scenarios.

However, since the Spectral Clustering procedure implicitly applies a non linear transformation from the *Input Space* (related the variations in pressure and flow) to the eigen-space spanned by the most relevant eigenvectors of the Laplacian Matrix, similarity computed in the Input Space does not guarantee that the association of the computed vectors to a specific cluster is correct. In order to improve the reliability of the localization process, a Support Vector Machine (SVM) classifier has been trained taking the variations in pressure and flow of each scenario as input and

the correspondent cluster provided by Spectral Clustering as target output (class label). Thus, the SVM classifier learns to approximate the non-linear mapping performed by Spectral Clustering and to estimate the most probable cluster which an actual vectors of variations in pressure and flow belongs to.

The Fig. 2 shows the overall workflow, presenting the mapping performed by the Spectral Clustering. The SVM permits to apply Spectral Clustering only to a more restricted, even if relevant, set of leakage scenarios, requiring a smaller scenarios graph, reducing complexity in memory and time, and guaranteeing to adopt a reliable approximation of Spectral Clustering to identify the correspondent scenarios cluster for any new vector of variations in flow and pressure.

2.4. Cost-effective sensors placement

Since variations in pressure and flow at the monitoring points are input both of Spectral Clustering and SVM classification, number and position of sensors affects overall performance of the approach proposed. Ideally, the greater the number of deployed sensors the higher is the quality of clustering and accuracy of the SVM classifier. However, deployment implies high costs for equipment and installation as well as useless redundancy of information. Optimal sensors placement has been recently addressed for both leakage detection/location [11, 12] and water quality issues [13, 14]. In this study, a solution for cost-effective sensor placement is proposed, aimed at identifying the best trade-off between high reliability in leakage localization (effectiveness) and costs for sensors. The solution uses, again, clustering on the dataset of leakage scenarios; in this case clustering (Partitioning Around Medoids, PAM) is applied on the columns of the dataset, that are variations in pressure and flow at each junction and each pipe, respectively. Clustering is performed separately for junctions and pipes sets, with the aim to group together, separately, those that are similar over the simulated leakage scenarios. Only the medoids of the clusters are selected as the most relevant monitoring points (that are a pressure meter in the case of junction medoid and a flow meter in the case of pipe medoid).

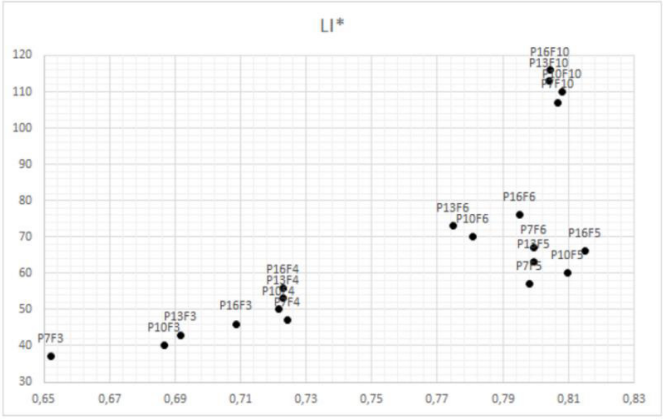

Fig. 3. Abbiategrasso: Costs for sensors (y-axis) versus LI* (x-axis).

3. Experimental results

In this section, the results are presented. The overall number of leakage scenarios that have been generated is 29800 and 3150 (both divided in 50% training and 50% test set), respectively for Abbiategrasso and Neptun, obtained by placing a leak, in turn, on each pipe, and varying its severity among 10 different values.

3.1. Results on cost-effective sensors placement

In Fig. 3 the results of all the sensors placements considered for Abbiategrasso are depicted: 7, 10, 13 or 16 pressure meters and 3, 4, 5 or 6 flow meters. Fig. 4 is related to the Neptun case study, where 7, 10 and 13 pressure meters and 1, 2 and 3 flow meters have been taken into account. *LI** has been considered as global index of localization reliability (effectiveness), while costs for sensors have been set 1 for a pressure meter and 10 for a flow meter.

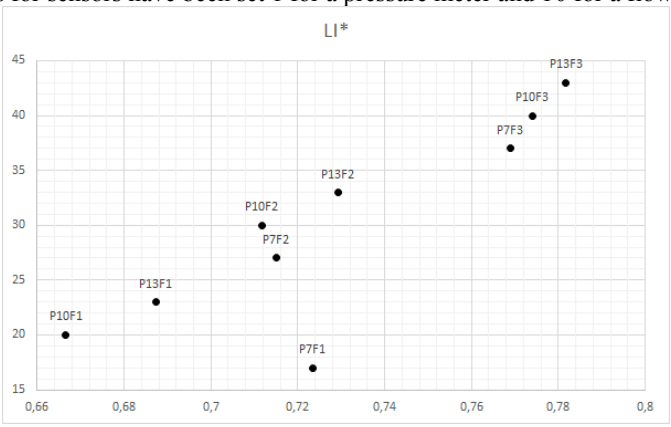

Fig. 4. Neptun: Costs for sensors (y-axis) versus LI* (x-axis)

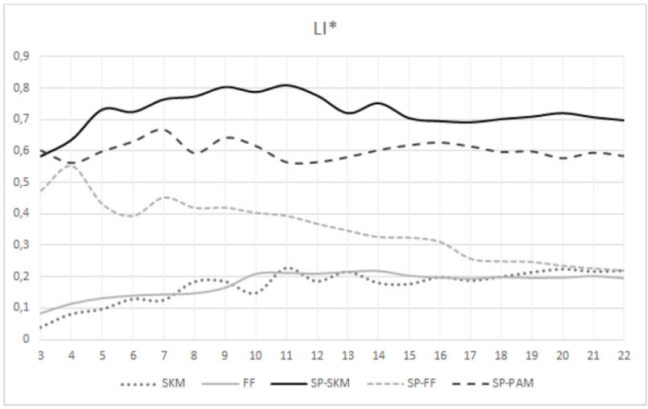

Fig. 5. Abbiategrasso: LI* depending on number of clusters (K), traditional versus network-based clustering algorithms: simple K-means (SKM), Farthest First (FF), Spectral Clustering (SP).

3.2. Results on Spectral Clustering

This section reports a comparison among three different implementations of Spectral Clustering (i.e., internally using: *a)* simple *K*-means, *b)* Farthest-First and *c)* PAM) and two "traditional", not-network-based clustering algorithms (i.e., simple *K*-means and Fartest-First). The Fig. 5 and Fig. 6 show the trend of *LI** with respect to the number of desired clusters for each algorithm. Taking into account the definition of *LI*, it is quite easy to understand that increasing *K* improves LI while reduces QL.

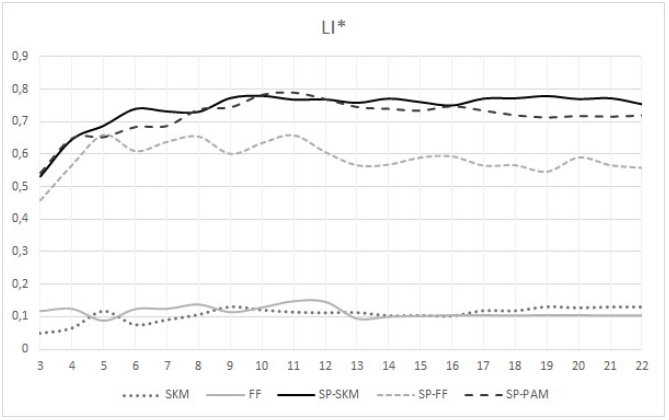

Fig. 6. Neptun: LI* depending on number of clusters (K), traditional versus network-based clustering algorithms: simple K-means (SKM), Farthest First (FF), Spectral Clustering (SP).

The best configuration selected in this paper is the algorithm Spectral Clustering with internally the simple *K*-means (S-SKM), by setting *K*=11 and *K*=12 respectively for Abbiategrasso e Neptun (and the first 3 smallest non-zero eigen-vectors in both cases). Spectral Clustering performances are clearly higher than those offered by clustering algorithms which are not-network-based.

3.3. Results on SVM-based leaky pipes identification

The implementation of C-SVM provided in the open-source Java-based suite WEKA (Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/) has been adopted. In particular, the Radial Basis Function (RBF) kernel has been adopted for non-linear mapping. Both *C* and the internal parameter γ of the kernel have been varied until accuracy, on 10 fold-cross validation, is no more improved. Accuracy is the percentage of vectors correctly associated to the cluster provided by the Spectral Clustering process; 10 fold-cross validation technique uses the entire dataset to train a model and test it, giving an estimation of the reliability in predicting the class label (i.e., cluster associated to new vectors of hydraulic variations, in this case). The best SVM configuration resulted setting C = 1 and γ = 1. The learned SVM classifier has been then validate on an independent test set, related to leakage scenarios obtained on values of severity different from those already adopted (i.e. new leaks). Neither Spectral Clustering or SVM training are performed on this test set; the vectors of pressure and flow variations, associated to a leak, are given as input to the learned SVM classifier, which provides an estimation of the cluster probably assigned by Spectral Clustering. If the leaky pipe associated to the specific vector of variations is in the set of distinct pipes associated to the predicted scenarios cluster a successful localization is counted for.

The following Fig. 7 and Fig. 8 summarize the performances related to the number of successful localizations, both on training (average = 97.99% and 97.21%, respectively for Abbiategrasso and Neptun) and independent test (average = 98.02% and 97.50%, respectively for Abbiategrasso and Neptun).

4. Conclusions

The approach presented in this paper aims at improving leakage localization in urban WDN through simulation of several leakage scenarios, Spectral Clustering and Support Vector Machine classification. A reliable relationship (reliability about 98%) between variations, in pressure and flow, and leak location has been identified and can be used to reduce time and costs for investigations and rehabilitation: when a leak is detected (e.g., with traditional methods, such as Minimum Night Flow analysis [3]) actual pressure and flow measurements are given as input to the SVM which provides the set of probably leaky pipes (cluster of the Spectral Clustering) associated to that variation. The

framework supports also cost-effective sensor placement. The overall approach has been validated on a real case study, the Abbiategrasso PMZ, in Milan, Italy, one of the two pilots of the European project ICeWater.

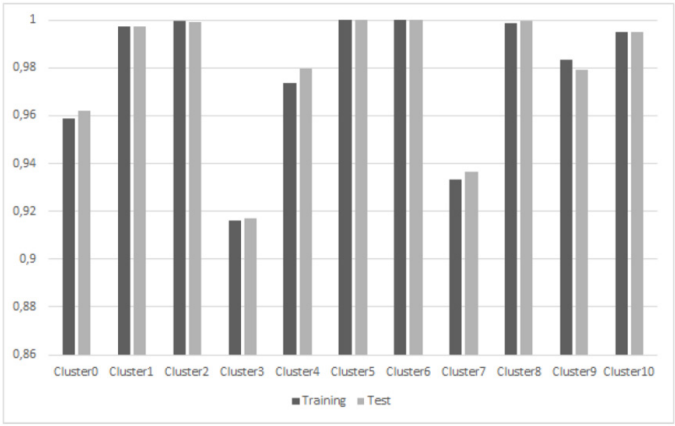

Fig. 7. Successful leakage localization, both on training and test set and for each cluster.

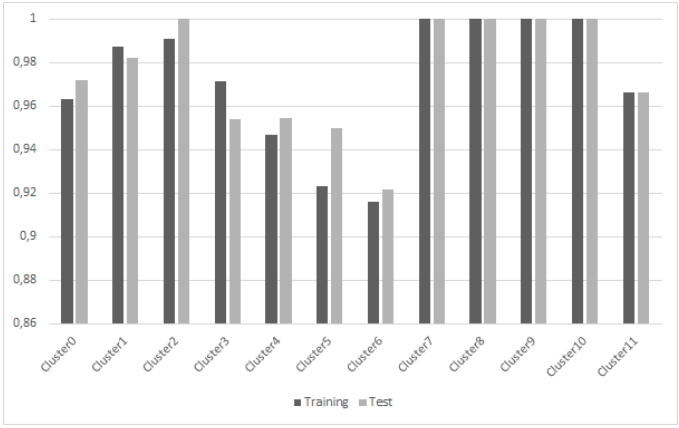

Fig. 8. Successful leakage localization, both on training and test set and for each cluster.

5. Acknowledgements

This work has been partially supported by the European Union ICeWater project – FP7-ICT 317624 (www.icewater-project.eu).

References

[1] H. Alegre, J.M. Baptista, E. Cabrera, F. Cubillo, P. Duarte, W. Hirner, W. Merkel, R. Parena, Performance Indicators for Water Supply Services, Second Edition, IWA Publishing (2006).

[2] R. Puust, Z. Kapelan, D.A. Savic, T. Koppel, A review of methods for leakage management in pipe networks, Urban Water Journal 7 (2010) 25-45.

[3] K. Behzadian, Z. Kapelan, D.A. Savic, A. Ardeshir, Stochastic sampling design using multi objective genetic algorithm and adaptive neural networks, Environmental Modeling and Software 24 (2009) 530–541.

[4] L. Xia, L. Guo-jin, Leak detection of municipal water supply network based on the cluster-analysis and fuzzy pattern recognition, International Conference on E-Product E-Service and E-Entertainment (ICEEE) 1 (2010) 7-9.
[5] A. Nasir, B.H. Soong, S. Ramachandran, Framework of WSN based human centric cyber physical in-pipe water monitoring system, Proc. 11th International Conference on Control, Automation, Robotics and Vision (2010) 1257-1261.
[6] W. Lijuan, Z. Hongwei, J. Hui, A Leak Detection Method Based on EPANET and Genetic Algorithm in Water Distribution Systems, Software Engineering and Knowledge Engineering: Theory and Practice – Advances in Intelligent and Soft Computing 14 (2012) 459-465.
[7] A. Candelieri, E. Messina, Sectorization and analytical leaks localizations in the H2OLeak project: Clustering-based services for supporting water distribution networks management, Environmental Engineering and Management Journal 11 (2012) 953-962.
[8] A. Candelieri, D. Conti, F. Archetti, A graph based analysis of leak localization in urban water networks, Proc. 12th International Conference on Computing and Control for the Water Industry, CCWI2013, (2013).
[9] A. Candelieri, F. Archetti, E. Messina, Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation, WIT Transactions on Ecology and the Environment 171 (2013) 107-117.
[10] J. Mashford, D. De Silva, S. Burn, D. Marney, Leak Detection in simulated water pipe networks using SVM, Applied Artificial Intelligence: An International Journal 26 (2012) 429-444.
[11] S.E. Christodoulou, A. Gagatsis, S. Xanthos, S. Kranioti, A. Agathokleous, M. Fragiadakis, Entropy-Based Sensor Placement Optimization for Waterloss Detection in Water Distribution Networks, Water Resour Management 27 (2013) 4443–4468.
[12] M.V. Casillas, V. Puig, L.E. Garza-Castanon, A. Rosich, Optimal Sensor Placement for Leak Location in Water Distribution Networks Using Genetic Algorithms, Sensors 13 (2013) 14984-15005.
[13] K. Klise, C. Phillips, R. Janke, Two-Tiered Sensor Placement for Large Water Distribution Network Models, Journal of Infrastructure Systems 19 (2013) 465–473.
[14] N.B. Chang, N.P. Pongsanone, A. Ernest, A rule-based decision support system for sensor deployment in small drinking water networks, Journal of Cleaner Production 29–30 (2012) 28-37.